ICLR

Perceptual Adversarial Robustness

Defense Against Unseen Threat Models

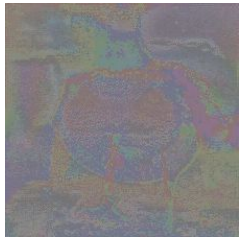
Cassidy Laidlaw, Sahil Singla, Soheil Feizi
University of Maryland

https://arxiv.org/abs/2006.12655
https://github.com/cassidylaidlaw/perceptual-advex

What's an adversarial example?

"American dipper"

+

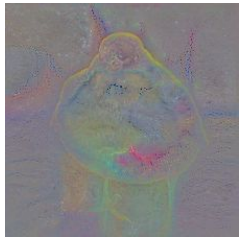


L∞

=

"academic gown"

+

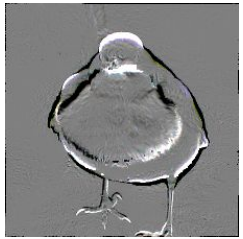


L_2

=

"Siberian husky"

+



StAdv

=

"hook"

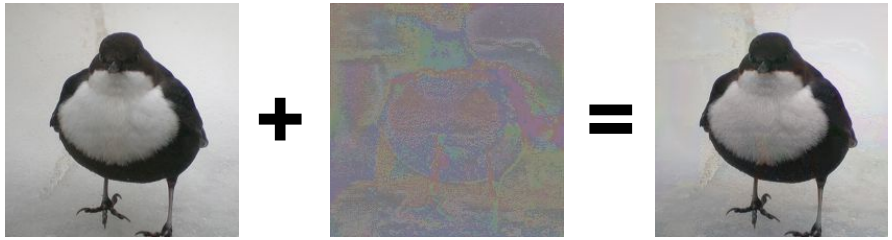
+

ReColorAdv

=

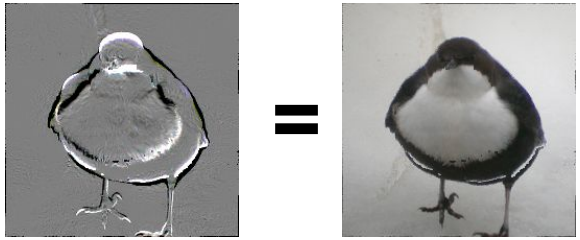
"red-backed sandpiper"

CIFAR-10 L∞ adversarial training

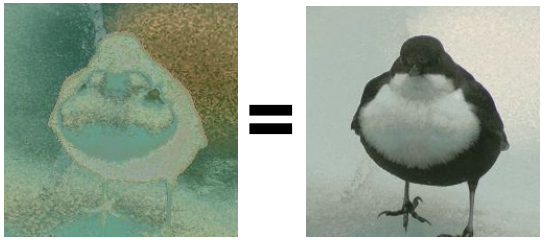


L∞ **49% accuracy**

L_2 **19% accuracy**



StAdv **5% accuracy**



ReColorAdv **55% accuracy**

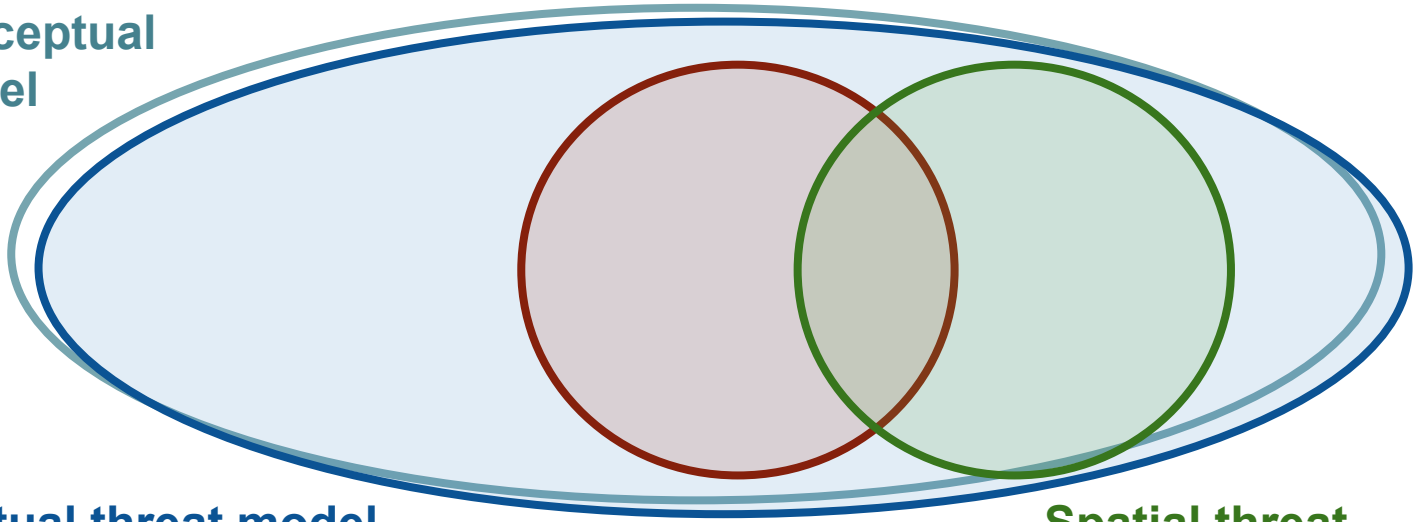
We need defenses that transfer to unseen perturbations!

Our contributions

- A unification of existing imperceptible adversarial threat models into a single threat model: the **perceptual adversarial threat model**

- A defense, **perceptual adversarial training (PAT)**, which uses this threat model to give robustness against unseen perturbation types

Key idea: most existing adversarial perturbation types are subsets of the set of all imperceptible perturbations.



Approximating the perceptual distance

For images x_1 and x_2, the **true perceptual distance** $d^*(x_1, x_2)$ is how different the images appear to humans.

The **perceptual adversarial threat model** includes all adversarial examples $\tilde{x}$ such that $d^*(x, \tilde{x}) \leq \varepsilon^*$. ε^* is the **perceptibility threshold**.

We approximate $d^*(x_1, x_2)$ with the LPIPS distance [1] $d(x_1, x_2)$ to form the **neural perceptual threat model (NPTM)**.

[1] Zhang et al. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. CVPR 2018.

LPIPS correlates well with human perception

Perceptual study: 21k human annotations across 7 adversarial threat models, each at 3 bounds.

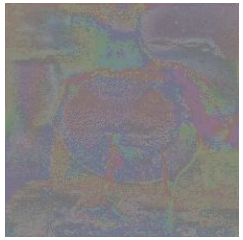
- L2 correlation with human perception: **0.88**
- LPIPS correlation with human perception: **0.94**

Perceptual adversarial examples

Accuracy of adversarial trained classifier against perceptual attacks on CIFAR-10:

0.3%

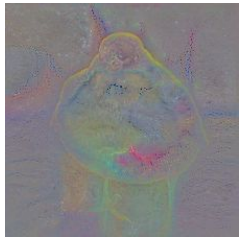
CIFAR-10 perceptual adversarial training (PAT)


+

=


L∞

~~49% accuracy~~

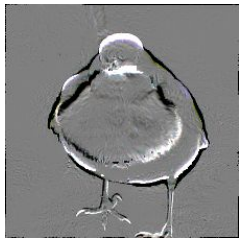
30% accuracy

+

=


L_2

~~19% accuracy~~

35% accuracy

+

=


StAdv

~~5% accuracy~~

46% accuracy

+

=


ReColorAdv

~~55% accuracy~~

71% accuracy